Dataset Methodology: Building a Comprehensive Account of Federal Arrests for Racially and Ethnically Motivated Violent Extremism and Targeted Violence in the US

Dataset Produced by the UNC Digital Propaganda Research Team:
Primary Dataset Team: A.A. Mattheis, M.B. Doty, A. Sin, M. Conley, C. Oh, S. Triana, and D. Antonetti; with support from: C. Dauber, M. Robinson, B. Ladd, and K. Papadopoulos

(This project was completed with the generous support of the Office of Undergraduate Research at the University of North Carolina at Chapel Hill)

About the UNC Dataset:

This dataset is comprised of manually collected and vetted information developed using open source, web-based data gathering. Our full team includes our primary research team comprised of one faculty member, the director of our media lab, and two graduate student specialists as well as six advanced undergraduate students from a variety of disciplines recruited from our related classes in terrorism, hate speech, linguistics and propaganda analysis. The entire team was needed due to the volume of data and the required parsing necessary to build the final dataset.

This dataset provides information about individuals arrested and charged for federal crimes in the US linked to 'domestic' terrorism stemming from white supremacist, nationalist, accelerationist, and male supremacist ideological cultures and groups. This dataset does not include charges at the state level, where many such crimes are charged depending on whether or not a case meets the requirements for federal charging. In some REMVE / TV cases, charges will occur at both the federal and state level (e.g. The Charleston AME Church shooter was charged at both levels). Given the complexities of searching 50 separate state court systems and the limitations of our team (size and time), we were unable to include state charges in the dataset.

The information includes the name and basic information of the person charged, whether the charge occurred within a multi-arrest operation, charges and case status, as well as case numbers, arrest, charging, and sentencing dates and details (where available). This information is all publicly available data under US law (thus does not breach privacy of individuals involved) that we have aggregated for ease of analysis. The file is in an excel file format as the data comprises more than 550 rows of total data. Data may be missing for cases because the case is ongoing and that data is not yet available or because that data was not available in any of the multiple open source sites for a particular case. The charges are coded into three subcategorizations to align with the new DHS Strategic Framework classification system: 1) 'REMVE' for crimes clearly linked to ideologies or groups, 2) 'TV – Hate Crimes' for cases that are not clearly linked to REMVE groups or ideology but which incurred hate crimes charges, and 3) 'TV – Other' to indicate cases that do not have clear REMVE links or did not incur hate crimes charges, such as crimes targeting government personnel or installations or attacks with male supremacist ideological links. These codes reflect our best understanding of the classifications given that the definition of 'targeted violence' remains unclear in the Framework document.

The dataset does not include individuals arrested or charged with crimes linked to REMVE ideologies such as members identified with the Black Hebrew Israelites who committed two

recent and closely timed attacks in Jersey City, NJ and Monsey, NY in December 2019. These groups and ideologies, generally categorized under "Black Separatism," are substantially less prevalent in the US than White supremacist groups and ideologies. For information on current and historical problems with US government designations of Black ideologies and groups the Marshall Project provides detailed information from a variety of reputable sources.

Our goal in building this dataset is to assist our team and other researchers in the systematic study of US federal charges and prosecutions associated with this growing threat. Moreover, the aggregation of this data allows for better understanding of charging patterns and distribution, case outcomes, and how such cases are treated across the various federal districts in the US. It is our hope that this dataset will be useful for researchers across a wide variety of disciplines and areas of interest and that developing it along with datasets for state criminal charges will become a community-wide effort going forward.

Building the Dataset:

The team wanted the students to be able to search without the added hurdles (time to learn, cost, etc.) of using PACER. Thus, this process describes a manual data-collection and verification process using open-source websites. As a starting point, the team decided to use the US Department of Justice (DOJ) website for initial searches because the US DOJ provides press releases for federal charges on their site. Prosecutors, however, use a wide variety of criminal charges when dealing with these defendants, so there is no simple way to search for them even among DOJ press releases. To address this issue, we employed manual vetting across all 93 federal districts. Manual vetting was the only way we could see to compose the dataset. This allowed the team to gather initial information about cases and charged individuals based on the press releases and use that information for secondary searches as needed to gather all the data we required.

While US DOJ press release searches are available from a single database, we needed to be able to distribute the workload amongst our researchers. The US DOJ manages its responsibilities (investigations, charging, and prosecutions among other duties) across the US through a network of 'districts' organized by geographic areas, each the responsibility of an Assistant US Attorney. Thus, we used the district system as a framework for organizing the data collection workload between the undergraduate researchers.

We initially distributed the districts between students by dividing the districts numerically (93 districts split among 6 students for an average number of 16 districts per student). Because the district lists are regional, we found that some districts are more likely than others to have higher volumes of REMVE / TV activity. We addressed this by redistributing some of the high-volume districts to rebalance the workload as needed. We met as a team weekly to review progress, work our way through problems, and come up with solutions to recurring issues. The detailed outline below encompasses the research process that we collaboratively developed.

1. Search key terms as defined by project parameters in the US DOJ press releases online. An initial broad search of the press releases can be narrowed by selecting DOJ

divisions in the search form under 'component' to select for likely divisions, especially the National Security Division and the Criminal Division.

2. Skim headlines on each district page to get a feel for 'typical' crimes in district. REMVE / TV often stands out.
3. Find perpetrator/defendant via press release.
    a. First, middle, and last names are essential. If press release does not provide the full name, Google the given press release name, district found, and crime to gather full name for data.
    b. If race is not provided, search other news outlets to find a potential mug shot.
    c. Select districts provide case numbers as well but if yours does not, use free legal resources such as https://www.courtlistener.com/ and https://www.govinfo.gov/.
4. Search for additional news coverage of the defendant by Googling name + district + crime, or by Googling "US vs. [insert full name]".
5. Review findings with team to verify dataset inclusions and exclusions as well as need for additional categorizations based on data.
6. Manually review each line item in the final list in order to resolve as many 'missing' data points (such as dates or sentence information) as possible.

Some cases are more likely to be included in news reports than others, so searching additional coverage sometimes yields a wealth of information and sometimes yields very little. It can, however, be an important step to finding out relationships between events and details that are less accessible when full charging documents or transcripts are unavailable. Searching with PACER, if your team has the requisite knowledge and can absorb the cost, would likely streamline this process and provide details unavailable even with multiple site searches as listed above.

Notes on Search Terminology:

Over the course of the project, students' use of terminology overlapped in some cases and diverged in others based on their individual relationship to the topic and experience with online research. A primary set of terms emerged among the student researchers that can be understood as 'correlational' terms. By correlational, we mean that these terms are common terms used to describe REMVE / TV groups, acts, and ideology in social contexts. There was a secondary set of terms that emerged that can be understood as 'associational' terms, or terms linked to specific groups. The last group of terms identified during the process can be described as 'situational' terms, or terms that describe criminal behaviors which are likely to be used by REMVE / TV actors. Each of the three sets of terms provide results with the correlational and associational terms providing the most closely aligned results and situational terms casting the widest net. We believe all three types of terminology are useful to capture charges given the inconsistency with which actors are linked in top-level documents with indicators of DT, REMVE, and TV. Examples of terminological groups used include:

"Correlational Terms"

White	Hate	Race / Racial / Racist	Jewish
White	Nazi	Anti-Semitic	Slur

Supremacist / Supremacy			
Supremacist / Supremacy	Aryan	Minority	Separatist
Terror	Extremist	Nationalist	Sexual

Note: Terms used are collapsed in list above where Boolean searching allows for using * at the stem to provide all permutations of the word.

"Associational Terms"

Brotherhood	Nation	KKK
Atomwaffen	Klan	Sovereign

"Situational Terms"

Firearms	Kingpin	Affiliate
Gang	Cult	Bomb
Conspiracy	Cyberstalking	Orientation

It is important to think about how many of each type of term to use and the level of parsing required to review the data returned by various types of terminology. Correlational and associational terms are more directly linked to the data, but may or may not be used within some documentation or may be used in various combinations depending on charges and how information about the charging is released on the region/district website. Situational terms will provide a broader set of data and require more follow up research to verify, but may catch items missed in the more targeted searches using correlational or associational terms. Moreover, these types of terms are useful because they describe regular types of charges related to criminal activity that supports REMVE groups such as drug and firearm trafficking and precursor acts related to TV events such as school and workplace attacks, and other violent events which are clearly targeted but which do not have a clear political or ideological connection.